

PEMBENTUKAN MODEL PROBIT BIVARIAT



SKRIPSI

Disusun Oleh :

Yudhi Candra

J2E 003 266

PROGRAM STUDI STATISTIKA

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS DIPONEGORO

SEMARANG

2009

PEMBENTUKAN MODEL PROBIT BIVARIAT

Disusun Oleh :

Yudhi Candra

J2E 003 266

Skripsi

**Sebagai salah satu syarat untuk memperoleh gelar Sarjana Sains
pada Program Studi Statistika**

PROGRAM STUDI STATISTIKA

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS DIPONEGORO

SEMARANG

2009

BAB I

PENDAHULUAN

1.1. LATAR BELAKANG

Model regresi merupakan komponen penting dalam beberapa analisis data dengan menggambarkan hubungan antara variabel respon dan satu atau beberapa variabel bebas. Pada umumnya analisis regresi digunakan untuk menganalisis data dengan variabel respon berupa data kuantitatif. Akan tetapi dalam kehidupan sehari-hari sering ditemui kasus dengan variabel responnya bersifat kualitatif, seperti keputusan memilih “ya” atau “tidak”. Untuk menyelesaikan kasus ini dapat digunakan model probit.

Model probit merupakan model non linier yang digunakan untuk menganalisis hubungan antara satu variabel respon dan beberapa variabel bebas, dengan variabel responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan sebuah karakteristik dan bernilai 0 untuk menyatakan ketidakberadaan sebuah karakteristik.

Model probit dengan satu variabel respon dapat dikembangkan menjadi model probit dengan menggunakan dua variabel respon, model ini disebut model probit bivariat. Model probit bivariat menggunakan dua variabel dikotomi sebagai variabel responnya, sedangkan variabel bebasnya dapat berupa variabel yang bersifat diskrit maupun variabel yang bersifat kontinu dan juga dapat berupa variabel kualitatif yaitu variabel nominal atau ordinal. Salah satu contoh kasus probit bivariat adalah penelitian yang dilakukan oleh Pindyck dan Rubinfeld pada

tahun 1973 di kota Troy, Michigan, Amerika Serikat tentang keputusan suatu keluarga apakah akan membelanjakan uangnya untuk membiayai salah seorang anaknya ke sekolah negeri atau akan membelanjakan uangnya untuk membayar pajak properti. Variabel yang diamati adalah pendapatan, pajak kekayaan, dan lama menetap dalam lingkungan tersebut. Akan diselidiki apakah variabel pendapatan, pajak kekayaan, dan lama menetap dalam lingkungan tersebut mempengaruhi keputusan seseorang lebih memilih membelanjakan uangnya untuk membiayai salah seorang anaknya ke sekolah negeri atau memilih membelanjakan uangnya untuk membayar pajak properti (Greene,1984). Karena dalam kasus tersebut data yang disajikan berupa data kualitatif, dan variabel responnya terdiri dari dua variabel maka untuk menganalisis hubungan antar variabelnya menggunakan pendekatan model probit bivariat.

Dalam penulisan tugas akhir ini akan dibahas mengenai bentuk model probit bivariat dan penaksiran parameternya menggunakan metode maksimum likelihood. Selanjutnya akan dibahas mengenai model probit bivariat untuk penerapan pada contoh kasus tentang faktor-faktor yang berpengaruh terhadap perolehan beasiswa nasional (*National Science Scholar*) dan penghargaan matematika tingkat universitas (*Moobery Math Award*) pada 300 orang mahasiswa UCLA, Amerika Serikat. Dalam pengolahan datanya akan digunakan alat bantu software Stata 10.

1.2. RUMUSAN MASALAH

Berdasarkan latar belakang tersebut akan dibahas mengenai pembentukan model probit bivariat dan penerapannya dalam suatu contoh kasus tentang faktor-faktor yang berpengaruh terhadap perolehan beasiswa nasional (*National Science Scholar*) dan penghargaan matematika tingkat universitas (*Moobery Math Award*) pada 300 orang mahasiswa UCLA, Amerika Serikat. Dari contoh kasus itu akan diperoleh suatu model yang menggambarkan pola hubungan antara dua variabel responnya dengan beberapa variabel bebasnya.

1.3. PEMBATASAN MASALAH

Pembatasan masalah pada Tugas Akhir ini adalah pembentukan model probit bivariat dan penentuan estimasi parameternya dengan metode iterasi Newton-Raphson.

1.4. TUJUAN PENULISAN

Tujuan dari penulisan Tugas Akhir ini adalah :

1. Menentukan estimasi parameter dan membentuk model probit bivariat.
2. Menguji korelasi masing-masing variabel respon dan menguji signifikansi parameter model secara keseluruhan maupun secara individu.
3. Mengaplikasikan model probit bivariat dalam suatu contoh kasus tentang faktor-faktor yang berpengaruh terhadap perolehan beasiswa nasional (*National Science Scholar*) dan penghargaan matematika

tingkat universitas (*Moobery Math Award*) pada 300 orang mahasiswa UCLA, Amerika Serikat, sehingga dapat diketahui hubungan antar variabelnya yang mampu menggambarkan dan menjelaskan permasalahan tersebut.

1.5. SISTEMATIKA PENULISAN

Sistematika penulisan tugas akhir ini adalah Bab I merupakan Pendahuluan yang berisi latar belakang, rumusan masalah, pembatasan masalah, dan sistematika penulisan. Bab II merupakan Teori Penunjang yang berisi konsep dasar penulisan, meliputi : probabilita, fungsi distribusi kumulatif, data kualitatif, distribusi normal bivariat, variabel dummy, metode maksimum likelihood, matriks Hessian, dan metode Newton Raphson. Bab III berisi pembahasan tentang Model Probit Bivariat yang berupa model umum, penaksiran parameter, uji signifikansi model, dan contoh penerapan. Bab IV merupakan kesimpulan dari bab-bab sebelumnya.

BAB II

TEORI PENUNJANG

2.1. Probabilita

Jika sebuah percobaan E mempunyai ruang sampel S dan sebuah kejadian A didefinisikan pada S , maka $P(A)$ adalah suatu angka riil yang disebut probabilita dari peristiwa A atau probabilita A dan fungsi $P(\cdot)$ mempunyai syarat-syarat sebagai berikut :

1. $0 \leq P(A) \leq 1$ untuk tiap kejadian A dari S
2. $P(\emptyset) = 0$
3. $P(S) = 1$

Bila suatu percobaan yang dapat menghasilkan n macam hasil yang berkemungkinan sama dan bila terdapat sebanyak $n(A)$ dari hasil yang berkaitan dengan kejadian A , maka probabilita A adalah :

$$P(A) = \frac{n(A)}{n} \quad (2.1.1)$$

Teorema 2.1

Bila A dan A' kejadian yang saling berkomplemen, maka :

$$P(A') = 1 - P(A)$$

Bukti : $S = A \cup A'$

$$A \cap A' = \emptyset$$

$$1 = P(S)$$

$$\begin{aligned}
 &= P(A \cup A') \\
 &= P(A) + P(A')
 \end{aligned}$$

sehingga

$$P(A') = 1 - P(A)$$

(Walpole, 1995)

2.2. Fungsi Distribusi Kumulatif

Definisi 2.1

Fungsi distribusi kumulatif atau *probabilitas kumulatif* sering disebut fungsi distribusi saja. Fungsi distribusi variabel random kontinu X yang dinotasikan $F_x(x) = P(X \leq x)$ untuk semua bilangan riil x , didefinisikan dengan :

$$F_x(x) = \int_{-\infty}^x f(t) dt \quad (2.2.1)$$

Sifat-sifat fungsi distribusi :

1. $\lim_{x \rightarrow \infty} F(x) = 1$
2. $\lim_{x \rightarrow -\infty} F(x) = 0$
3. Fungsi tersebut tidak turun, yaitu jika $b \geq a$ maka $F(b) \geq F(a)$
4. Fungsi tersebut kontinu dari kanan, yaitu untuk seluruh x dan $\delta > 0$

$$\lim_{\delta \rightarrow 0} [F(x + \delta) - F(x)] = 0$$

5. $P(X > x) = 1 - F(x)$ atau $P(X > x) = 1 - P(X \leq x)$

(Hines dan Montgomery, 1990)

2.3. Data Kualitatif

Data kualitatif adalah data yang bukan merupakan angka yang skala pengukurannya terdiri dari seperangkat kategori. Data kualitatif terbagi menjadi dua yaitu data nominal dan ordinal.

a. Data Nominal

Data nominal adalah data yang berasal dari pengelompokan peristiwa berdasarkan kategori tertentu yang perbedaannya hanyalah menunjukkan perbedaan kualitatif.

Data ini tidak menggambarkan kedudukan objek atau kategori tersebut terhadap objek atau kategori lainnya tetapi hanya sekedar label atau kode saja. Data ini hanya mengelompokkan objek atau kategori ke dalam kelompok tertentu. Data ini memiliki dua ciri yaitu :

- Kategori data bersifat saling lepas (satu objek hanya masuk pada satu kelompok saja).
- Kategori data tidak disusun secara logis.

Contoh dari data nominal antara lain jenis kelamin (1 untuk pria, 0 untuk wanita), agama (1 untuk Islam, 2 untuk Kristen, 3 untuk Katolik, 4 untuk Budha, dan 5 untuk Hindu).

b. Data Ordinal

Data ordinal adalah data yang berasal dari objek atau kategori yang disusun menurut besarnya, dari tingkat terendah ke tingkat tertinggi atau sebaliknya, dengan jarak atau rentang yang tidak harus sama.

Data ini memiliki ciri seperti pada data nominal ditambah satu ciri lagi, yaitu kategori data dapat disusun berdasarkan urutan logis dan sesuai dengan besarnya karakteristik yang dimiliki.

Contoh dari data ordinal antara lain jenjang kepangkatan pada militer (jendral, kolonel, kapten, sersan), kelas sosial masyarakat (atas, menengah, kebawah).

(Hasan, 2004)

2.4. Distribusi Normal Bivariat

Misalkan sebuah fungsi ditunjukkan oleh :

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-q/2}, \quad -\infty < x < \infty, -\infty < y < \infty \quad (2.4.1)$$

Dengan $\sigma_1 > 0$, $\sigma_2 > 0$, dan $-1 < \rho < 1$,

$$q = \frac{1}{1-\rho^2} \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] \quad (2.4.2)$$

Dalam fungsi tersebut ditunjukkan :

- $f(x, y)$ adalah fungsi densitas bersama normal bivariat.
- $X \sim N(\mu_1, \sigma_1^2)$ dan $Y \sim N(\mu_2, \sigma_2^2)$.
- ρ adalah koefisien korelasi dari X dan Y .
- Variabel random X dan Y dikatakan memiliki distribusi normal bivariat.

Fungsi $f(x, y)$ yang merupakan fungsi densitas bersama dapat dilihat sebagai berikut.

Didefinisikan $f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$

Dari (2.4.2) diperoleh :

$$\begin{aligned} (1 - \rho^2)q &= \left[\left(\frac{y - \mu_2}{\sigma_2} \right) - \rho \left(\frac{x - \mu_1}{\sigma_1} \right) \right]^2 + (1 - \rho^2) \left(\frac{x - \mu_1}{\sigma_1} \right)^2 \\ &= \left(\frac{y - b}{\sigma_2} \right)^2 + (1 - \rho^2) \left(\frac{x - \mu_1}{\sigma_1} \right)^2 \end{aligned}$$

dimana $b = \mu_2 + \rho(\sigma_2 / \sigma_1)(x - \mu_1)$.

Sehingga :

$$f_1(x) = \frac{\exp \left[-(x - \mu_1)^2 / 2\sigma_1^2 \right]}{\sigma_1 \sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\exp \left\{ -(y - b)^2 / [2\sigma_2^2(1 - \rho^2)] \right\}}{\sigma_2 \sqrt{1 - \rho^2} \sqrt{2\pi}} dy \quad (2.4.3)$$

Dari integral yang terdapat pada $f_1(x)$ integrannya merupakan fungsi distribusi kumulatif normal dengan rata-rata b dan variansi $\sigma_2^2(1 - \rho^2)$, sehingga integral tersebut sama dengan 1 dan $f_1(x)$ menjadi :

$$f_1(x) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left[-\frac{(x - \mu_1)^2}{2\sigma_1^2} \right], \quad -\infty < x < \infty \quad (2.4.4)$$

Fungsi $f(x, y)$ merupakan fungsi densitas bersama dari dua variabel random kontinu X dan Y . Maka fungsi $f_1(x)$ adalah fungsi distribusi marginal untuk X , dan X berdistribusi $N(\mu_1, \sigma_1^2)$. Sehingga dapat dianalogkan Y berdistribusi $N(\mu_2, \sigma_2^2)$.

Dari hal itu, dapat dikembangkan bahwa :

$$f(x, y) = f_1(x) \left(\frac{1}{\sigma_2 \sqrt{1 - \rho^2} \sqrt{2\pi}} \exp \left[-\frac{(y - b)^2}{2\sigma_2^2(1 - \rho^2)} \right] \right) \quad (2.4.5)$$

dimana $b = \mu_2 + \rho(\sigma_2 / \sigma_1)(x - \mu_1)$.

Sehingga fungsi densitas untuk Y diberikan oleh $X = x$ adalah normal dengan rata-rata $b = \mu_2 + \rho(\sigma_2 / \sigma_1)(x - \mu_1)$ dan variansi $\sigma_2^2(1 - \rho^2)$. Maka dengan distribusi normal bivariat, rata-rata dari Y yang diberikan oleh $X = x$ adalah linear di x dan :

$$E(Y|x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1) \quad (2.4.6)$$

Sehingga dapat ditunjukkan dari distribusi oleh X , yang diberikan oleh $Y = y$ adalah distribusi normal $N\left[\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right]$.

Contoh 2.1

Diasumsikan bahwa populasi dari pasangan pernikahan yang memuat tinggi badan suami (X_1) dan tinggi badan istri (X_2) mengikuti distribusi normal bivariat dengan parameter $\mu_1=5,8$ kaki, $\mu_2=5,3$ kaki, $\sigma_1=\sigma_2=0,2$ kaki, dan $\rho=0,6$. fungsi distribusi kumulatif X_2 diberikan oleh $x_1=6,3$ adalah normal dengan mean $5,3+(0,6)(6,3-5,8)=5,6$ dan standar deviasi $(0,2)\sqrt{(1-0,36)}=0,16$.

Berdasarkan bahwa tinggi badan suami adalah 6,3 kaki, probabilitas bahwa seorang istri memiliki tinggi badan diantara 5,28 dan 5,92 kaki adalah :

$$P(5,28 < X_2 < 5,92 | x_1 = 6,3) = N(2) - N(-2) = 0,954$$

Teorema 2.2

Misalkan X dan Y mengikuti distribusi normal bivariat dengan rata-rata μ_1 dan μ_2 , variansi yang positif σ_1^2 dan σ_2^2 , dan koefisien korelasi ρ . Maka X dan Y secara stokastik independen jika dan hanya jika $\rho = 0$.

(Hogg dan Craig, 1995)

2.5. Variabel Dummy

Variabel dummy disebut juga variabel boneka, variabel indikator, variabel biner (2 angka), variabel bersifat kategori, dan variabel kualitatif. Pada umumnya variabel dummy untuk dua kategori diberi kode 0 dan 1.

Ciri model regresi dengan variabel dummy adalah sebagai berikut :

1. Jika suatu variabel kualitatif mempunyai J kategori, maka ada $J - 1$ variabel dummy.
2. Penetapan nilai 0 dan 1 untuk dua kategori, seperti pria dan wanita adalah tanpa suatu dasar atau bukan merupakan hal yang mutlak dalam variabel dummy mengambil nilai 0 dan 1 sebagai kode dari kategori tersebut.
3. Kelompok, kategori, atau klasifikasi yang diberi nol seringkali disebut sebagai kategori dasar, kategori kontrol, atau kategori perbandingan. Dengan kata lain merupakan perbandingan yang dibuat dalam kategori tersebut.

(Gujarati dan Zain, 1991)

2.6. Metode Maksimum Likelihood

Metode maksimum likelihood merupakan salah satu cara untuk mengestimasi parameter yang tidak diketahui. Prosedur estimasi maksimum likelihood menguji apakah estimasi maksimum yang tidak diketahui dari fungsi likelihood suatu sampel nilainya sudah memaksimumkan fungsi likelihood.

Misalkan X_1, X_2, \dots, X_N adalah variabel random dari populasi dengan distribusi fungsi $f(x, \theta)$, dengan θ adalah parameter yang tidak diketahui. Maka fungsi likelihood sampel tersebut adalah :

$$L(x_1, x_2, \dots, x_N; \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \dots f(x_N, \theta) \quad (2.6.1)$$

Bila fungsi likelihood terdeferensialkan ke- θ , maka estimasi maksimum likelihood dapat diperoleh melalui persamaan berikut :

$$(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_N) \rightarrow \frac{\partial L(x_1, x_2, \dots, x_N; \theta)}{\partial \theta_i} = 0 \quad (2.6.2)$$

$$i = 1, 2, \dots, N$$

Dalam banyak kasus, penggunaan deferensiasi akan lebih mudah bekerja pada logaritma natural dari $L(x_1, x_2, \dots, x_N; \theta)$, yaitu :

$$K(x_1, x_2, \dots, x_N; \theta) = \ln L(x_1, x_2, \dots, x_N; \theta) \quad (2.6.3)$$

Langkah-langkah untuk menentukan estimasi maksimum likelihood dari θ_i adalah :

1. Menentukan fungsi likelihood

$$L(x_1, x_2, \dots, x_N; \theta_i) = f(x_1, \theta_i) \cdot f(x_2, \theta_i) \dots f(x_N, \theta_i)$$

2. Membentuk logaritma natural likelihood :

$$K(x_1, x_2, \dots, x_N; \theta_i) = \ln L(x_1, x_2, \dots, x_N; \theta_i)$$

3. Membentuk persamaan likelihood dan menyelesaikan :

$$\frac{\partial K(x_1, x_2, \dots, x_N; \theta_i)}{\partial \theta_i} = 0$$

4. Didapat estimasi maksimum likelihood dari θ_i , yaitu $\hat{\theta}_i$.

5. Membuktikan bahwa $\hat{\theta}_i$ benar-benar memaksimumkan

$$L(x_1, x_2, \dots, x_N; \theta_i) \text{ dengan } \left. \frac{\partial^2 L}{\partial \hat{\theta}_i^2} \right|_{\theta_i = \hat{\theta}_i} < 0 \text{ dimana } i = 1, 2, \dots, N.$$

2.7. Matriks Hessian

Matriks Hessian adalah matriks bujur sangkar dari turunan parsial orde kedua. Misal didefinisikan fungsi real f sebagai berikut : $f(x_1, x_2, \dots, x_n)$. Jika turunan parsial orde kedua untuk semua f terdefinisi, maka matriks Hessian dari fungsi f adalah :

$$\mathbf{H}(f(x)) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

dengan $x = (x_1, x_2, \dots, x_n)$.

(Mital, K.V., 1987)

2.8. Metode Newton-Raphson

Metode Newton-Raphson adalah suatu metode untuk menyelesaikan sistem persamaan yang tidak linier (Agresti, 1990). Metode Newton-Raphson dapat dikembangkan dari perluasan deret Taylor, yang dapat dinyatakan sebagai :

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2} f''(x_n) + \dots \quad (2.8.1)$$

Untuk $n = 0, 1, 2, \dots$

Suku-suku orde kedua dari perluasan deret Taylor di sekitar x_n adalah :

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2} f''(x_n) \quad (2.8.2)$$

Jika x terdiri dari x_1, x_2, \dots, x_m dan $f(x_1, x_2, \dots, x_m)$ dapat ditulis $f(x)$,

$$f'(x) = T(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_m} \end{bmatrix} \text{ serta } f''(x) = \mathbf{H}(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \frac{\partial^2 f}{\partial x_m \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}$$

maka persamaan (2.8.2) dapat ditulis dengan :

$$f(x) = f(x_n) + (x - x_n)T(x_n) + \frac{(x - x_n)^2}{2} \mathbf{H}(x_n) \quad (2.8.3)$$

Turunan dari persamaan (2.8.3) terhadap x adalah :

$$T(x) = T(x_n) + (x - x_n)\mathbf{H}(x_n) \quad (2.8.4)$$

Jika $T(x) = 0$, maka akan diperoleh :

$$T(x_n) + (x - x_n)\mathbf{H}(x_n) = 0 \quad (2.8.5)$$

Aproksimasi yang baik dari x_n adalah x_{n+1} , maka persamaan (2.8.5) dapat

ditulis :

$$T(x_n) + (x_{n+1} - x_n)\mathbf{H}(x_n) = 0 \quad (2.8.6)$$

Dengan menyelesaikan persamaan (2.8.6), diperoleh suatu iterasi berikut :

$$\begin{aligned}
 T(x_n) + (x_{n+1} - x_n)\mathbf{H}(x_n) &= 0 \\
 (x_{n+1} - x_n)\mathbf{H}(x_n) &= -T(x_n) \\
 (x_{n+1} - x_n) &= -\mathbf{H}(x_n)^{-1}T(x_n) \\
 x_{n+1} &= x_n - \mathbf{H}(x_n)^{-1}T(x_n)
 \end{aligned} \tag{2.8.7}$$

(Atkinson, 1989)